CLAIMS

What is claimed is:

1. An apparatus for distributing a service request to one of a plurality of servers, each one of the plurality of servers being associated with a Domain Name System host name and each one of the plurality of servers having a unique IP address, comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

receiving a server request, the server request being a Domain Name System host name query including the Domain Name System host name;

incrementing a total number of server requests processed by the plurality of servers;

maintaining a number of server requests distributed to each one of the plurality of servers;

selecting one of the plurality of servers using a portion metric assigned to each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers; and

providing an IP address associated with the selected one of the plurality of servers.

2. The apparatus as recited in claim 1, wherein selecting one of the plurality of servers further includes:

computing a metric value for each one of the plurality of servers using the number of server requests distributed to each one of the plurality of servers, the portion metric assigned to each one of the plurality of servers, and the total number of server requests processed; and

comparing the metric value for each one of the plurality of servers to determine a selected server.

3. The apparatus as recited in claim 2, wherein maintaining the number of server requests distributed to each one of the plurality of servers further includes:

increasing the number of server requests distributed for the selected server.

4. The apparatus as recited in claim 2, wherein the selected server has a lowest metric value.

5. The apparatus as recited in claim 2, wherein the selected server has a highest metric value.

6. The apparatus as recited in claim 3, wherein increasing the number of server requests distributed is performed in response to distributing the server request.

7. The apparatus as recited in claim 1, wherein incrementing the total number of server requests further includes:

incrementing the total number of server requests in response to receiving the server request.

8. The apparatus as recited in claim 1, wherein incrementing the total number of server requests further includes:

incrementing the total number of server requests in response to distributing the server request.

9. The apparatus as recited in claim 1, wherein at least one of the processor and the memory are further adapted for:

if more than one server is selected, applying an alternate metric to a set of the plurality of servers to obtain a selected server.

10. The apparatus as recited in claim 2, wherein at least one of the processor and the memory are further adapted for:

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers.

11. The apparatus as recited in claim 10, wherein comparing the metric value for each one of the plurality of servers further includes:

defining a set of the plurality of servers having metric values within the tolerance range and including the selected one of the plurality of servers; and

applying an alternate metric to the set of the plurality of servers to obtain a selected one of the plurality of servers when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected one of the plurality of servers is the set of the plurality of servers.

12. The apparatus as recited in claim 10, wherein comparing the metric value for each one of the plurality of servers further includes:

obtaining a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range; and

applying an alternate metric to the set of the plurality of servers to obtain a selected one of the plurality of servers when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected one of the plurality of servers is the set of the plurality of servers.

13. The apparatus as recited in claim 2, wherein computing the metric value for each one of the plurality of servers further includes:

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

calculating a server request percentage for each one of the plurality of servers using the number of server requests distributed to each one of the plurality of servers, the total portion metric, and the total number of server requests received; and

determining a metric value for each one of the plurality of servers using the server request percentage distributed to each one of the plurality of servers and the portion metric assigned to each one of the plurality of servers.

14. The apparatus as recited in claim 13, wherein calculating the server request percentage for each one of the plurality of servers further includes:

calculating a server request percentage for one of the plurality of servers using the number of server requests for the one of the plurality of servers, the total portion metric, and the total number of server requests received.

15. The apparatus as recited in claim 14, wherein calculating the server request percentage for the one of the plurality of servers includes:

multiplying the number of server requests distributed to the one of the plurality of servers and the total portion metric to obtain a product; and

dividing the product by the total number of server requests received to obtain a server request percentage for the one of the plurality of servers.

16. The apparatus as recited in claim 13, wherein determining the metric value for each one of the plurality of servers further includes:

determining a metric value for one of the plurality of servers using the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers.

17. The apparatus as recited in claim 16, wherein determining the metric value for the one of the plurality of servers further includes:

subtracting the portion metric for the one of the plurality of servers from the server request percentage for the one of the plurality of servers to obtain a metric value for the one of the plurality of servers.

18. The apparatus as recited in claim 16, wherein determining the metric value for the one of the plurality of servers further includes:

subtracting the server request percentage for the one of the plurality of servers from the portion metric for one of the plurality of servers to obtain a metric value for the one of the plurality of servers.

19. The apparatus as recited in claim 15, wherein dividing the product by the total number of server requests further includes:

if the total number of server requests is equal to zero, initializing the server request percentage for the one of the plurality of servers to a value of zero; and

if the total number of server requests is greater than zero, dividing the product by the total number of server requests received to obtain the server request percentage for the one of the plurality of servers.

20. The apparatus as recited in claim 13, wherein calculating the server request percentage for each one of the plurality of servers further includes:

assigning a value of zero to the server request percentage for each one of the plurality of servers when the total number of server requests is equal to zero.

21. The apparatus as recited in claim 1, wherein at least one of the processor and the memory are further adapted for:

initializing the total number of server requests and the number of server requests distributed to each one of the plurality of servers to a constant.

22. The apparatus as recited in claim 21, wherein the constant is zero.

23. The apparatus as recited in claim 1, wherein incrementing the total number of server requests further includes:

adding the number of server requests distributed to each one of the plurality of servers to obtain the total number of server requests.

24. An apparatus for distributing a service request, comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

determining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a selected server;

applying an alternate metric to the selected server when the selected server includes more than one server; and

providing an IP address associated with the selected server.

25. The apparatus as recited in claim 24, wherein the alternate metric is a distance metric.

26. The apparatus as recited in claim 24, wherein at least one of the processor and the memory further comprise:

turning a plurality of metrics on, the plurality of metrics including the alternate metric.

27. The apparatus as recited in claim 24, wherein at least one of the processor and the memory further comprise:

specifying an order in which each of the plurality of metrics is considered.

28. The apparatus as recited in claim 24, wherein at least one of the processor and the memory further comprise:

specifying a priority for each one of the plurality of metrics.

29. The apparatus as recited in claim 28, wherein specifying a priority further includes:

assigning a weight indicating a metric priority to one of the plurality of metrics.

30. An apparatus for distributing a service request, the apparatus comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and

providing an IP address associated with the selected server.

31. An apparatus for detecting load imbalance within a distributed system, comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the

number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

detecting load imbalance when the set of the plurality of servers includes only one server; and

applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers.

32. The apparatus as recited in claim 31, wherein at least one of the processor and the memory are further adapted for:

generating a report in response to the detection of load imbalance, the report including information designed to assist in correcting load imbalance.

33. The apparatus as recited in claim 32, wherein the report includes the number of server requests distributed to each one of the plurality of servers.

34. The apparatus as recited in claim 31, wherein at least one of the processor and the memory are further adapted for:

reassigning portion metrics for selected ones of the plurality of servers in response to the detection of load imbalance.

35. The apparatus as recited in claim 31, wherein at least one of the processor and the memory are further adapted for:

reconfiguring the tolerance range in response to the detection of load imbalance.

36. An apparatus for distributing a service request, comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

accepting an HTTP connection;

incrementing a total number of server requests processed by the plurality of servers;

maintaining a number of server requests distributed to each one of the plurality of servers;

selecting one of the plurality of servers using the portion metric associated with each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests; and

sending an HTTP code redirect.

37. An apparatus for distributing a service request , comprising:

a processor; and

a memory, at least one of the processor and the memory being adapted for:

determining a metric value associated with each one of a plurality of servers;

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and

providing an IP address associated with the selected server.

38. A method for distributing a service request to one of a plurality of servers, each one of the plurality of servers being associated with a Domain Name System host name and each one of the plurality of servers having a unique IP address, comprising:

receiving a server request, the server request being a Domain Name System host name query including the Domain Name System host name;

incrementing a total number of server requests processed by the plurality of servers;

maintaining a number of server requests distributed to each one of the plurality of servers;

selecting one of the plurality of servers using a portion metric assigned to each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers; and

providing an IP address associated with the selected one of the plurality of servers.

39. A method for distributing a service request, the method comprising:

determining a portion metric assigned to each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the

number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a selected server;

applying an alternate metric to the selected server when the selected server includes more than one server; and

providing an IP address associated with the selected server.

40. A method for distributing a service request, comprising:

ascertaining a portion metric assigned to each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the

number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and

providing an IP address associated with the selected server.

41. A method for detecting load imbalance within a distributed system, comprising:

ascertaining a portion metric assigned to each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

maintaining a number of server requests distributed to each one of the plurality of servers;

receiving a server request;

incrementing a total number of server requests processed by the plurality of servers;

computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

detecting load imbalance when the set of the plurality of servers includes only one server; and

applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers.

42. A method for distributing a service request, comprising:

ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

accepting an HTTP connection;

incrementing a total number of server requests processed by the plurality of servers;

maintaining a number of server requests distributed to each one of the plurality of servers;

selecting one of the plurality of servers using the portion metric associated with each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests; and

sending an HTTP code redirect.

43. An apparatus for distributing a service request to one of a plurality of servers, each one of the plurality of servers being associated with a Domain Name System host name and each one of the plurality of servers having a unique IP address, comprising:

means for receiving a server request, the server request being a Domain Name System host name query including the Domain Name System host name;

means for incrementing a total number of server requests processed by the plurality of servers;

means for maintaining a number of server requests distributed to each one of the plurality of servers;

means for selecting one of the plurality of servers using a portion metric assigned to each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total

number of server requests, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers; and

means for providing an IP address associated with the selected one of the plurality of servers.

44. A computer readable medium for distributing a service request to one of a plurality of servers, each one of the plurality of servers being associated with a Domain Name System host name and each one of the plurality of servers having a unique IP address, the computer readable medium storing thereon the following instructions:

instructions for receiving a server request, the server request being a Domain Name System host name query including the Domain Name System host name;

instructions for incrementing a total number of server requests processed by the plurality of servers;

instructions for maintaining a number of server requests distributed to each one of the plurality of servers;

instructions for selecting one of the plurality of servers using a portion metric assigned to each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers; and

instructions for providing an IP address associated with the selected one of the plurality of servers.

45. An apparatus for distributing a service request, comprising:

means for determining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

means for adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

means for maintaining a number of server requests distributed to each one of the plurality of servers;

means for receiving a server request;

means for incrementing a total number of server requests processed by the plurality of servers;

means for computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

means for calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

means for comparing the metric value for each one of the plurality of servers to obtain a selected server;

means for applying an alternate metric to the selected server when the selected server includes more than one server; and

means for providing an IP address associated with the selected server.

46. A computer readable medium for distributing a service request, the computer readable medium storing thereon the following instructions:

instructions for determining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

instructions for adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

instructions for maintaining a number of server requests distributed to each one of the plurality of servers;

instructions for receiving a server request;

instructions for incrementing a total number of server requests processed by the plurality of servers;

instructions for computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

instructions for calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

instructions for comparing the metric value for each one of the plurality of servers to obtain a selected server;

instructions for applying an alternate metric to the selected server when the selected server includes more than one server; and

instructions for providing an IP address associated with the selected server.

47. An apparatus for distributing a service request, the apparatus comprising:

means for ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

means for adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

means for maintaining a number of server requests distributed to each one of the plurality of servers;

means for receiving a server request;

means for incrementing a total number of server requests processed by the plurality of servers;

means for computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of servers and the total portion metric divided by the total number of server requests received;

means for calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

means for ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

means for comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

means for applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and

means for providing an IP address associated with the selected server.

48. An apparatus for detecting load imbalance within a distributed system, comprising:

means for ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

means for adding the portion metric for each one of the plurality of servers to obtain a total portion metric;

means for maintaining a number of server requests distributed to each one of the plurality of servers;

means for receiving a server request;

means for incrementing a total number of server requests processed by the plurality of servers;

means for computing a server request percentage for each one of the plurality of servers, the server request percentage for one of the plurality of servers being a product of the number of server requests distributed to the one of the plurality of

servers and the total portion metric divided by the total number of server requests received;

means for calculating a metric value for each one of the plurality of servers, the metric value for one of the plurality of servers being defined by the server request percentage for the one of the plurality of servers and the portion metric for the one of the plurality of servers;

means for ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

means for comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

means for detecting load imbalance when the set of the plurality of servers includes only one server; and

means for applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers.

49. An apparatus for distributing a service request, comprising:

means for ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

means for accepting an HTTP connection;

means for incrementing a total number of server requests processed by the plurality of servers;

means for maintaining a number of server requests distributed to each one of the plurality of servers;

means for selecting one of the plurality of servers using the portion metric associated with each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests; and

means for sending an HTTP code redirect.

50. A computer readable medium for distributing a service request, the computer readable medium storing thereon the following instructions:

instructions for ascertaining a portion metric associated with each one of a plurality of servers, the portion metric designating a portion of total server requests to be allocated to the one of the plurality of servers;

instructions for accepting an HTTP connection;

instructions for incrementing a total number of server requests processed by the plurality of servers;

instructions for maintaining a number of server requests distributed to each one of the plurality of servers;

instructions for selecting one of the plurality of servers using the portion metric associated with each one of the plurality of servers, the number of server requests distributed to each one of the plurality of servers, and the total number of server requests; and

instructions for sending an HTTP code redirect.

51. An apparatus for distributing a service request , comprising:

means for determining a metric value associated with each one of a plurality of servers;

means for ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

means for comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

means for applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and

means for providing an IP address associated with the selected server.

52. A computer readable medium for distributing a service request , the computer readable medium storing thereon the following instructions:

instructions for determining a metric value associated with each one of a plurality of servers;

instructions for ascertaining a tolerance range used in comparing the metric value for each one of the plurality of servers;

instructions for comparing the metric value for each one of the plurality of servers to obtain a set of the plurality of servers, each one of the set of the plurality of servers having a metric value within the tolerance range;

instructions for applying an alternate metric to the set of the plurality of servers to obtain a selected server when the set of the plurality of servers includes more than one server, and otherwise establishing that the selected server is the obtained set of the plurality of servers; and